Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human-influences

Jamal Zaherpour (lgxjz1@nottingham.ac.uk), Simon N. Gosling, Nick Mount, Dieter Gerten, Hannes M. Schmied, Ingjerd Haddeland, Jacob Schewe, Junguo Liu, Guoyong Leng, Lukas Gudmundsson, Naota Hanasaki, Rutger Dankers, Stephanie Eisner, Taikan Oki, Ted Veldkamp, Yadu Pokhrel, Yoshimitsu Masaki, Yusuke Satoh, Yoshihide Wada

University of Nottingham
UK | CHINA | MALAYSIA

1. Summary We evaluate simulations of monthly runoff from six GHMs that participated in ISIMIP2a, across 40 catchments in 8 hydrobelts globally. The performance of each individual model and the ensemble mean, EM, in replicating observed mean and extreme runoff under human-influenced conditions (water withdrawals and dams) is assessed. Application of a novel integrated evaluation metric shows that generally, when assessing the timeseries of runoff, the models perform better in the wetter, equatorial and northern hydrobelts, than in drier, southern hydrobelts. When model outputs are temporally aggregated to assess mean annual and extreme runoff, the models perform better than when their timeseries are evaluated. However, the general trend in the majority of models is towards the overestimation of mean annual and extreme runoff. For all hydrological indicators, the EM of the models generally fails to perform better than any individual model – a finding that challenges the commonly held perception that the EM delivers superior performance over individual models.

2. Methods

2.1. Study area and data sources

- 40 large catchments (area >= 100,000 km^2 and observed data >= 25 years) across 8 hydrobelts (Meybeck et al. 2013).
- Monthly observed runoff data for 40 years (1971 – 2010) from the GRDC.
- Simulated runoff from 6 ISIMIP2a GHMs, openly available from the ESGF.

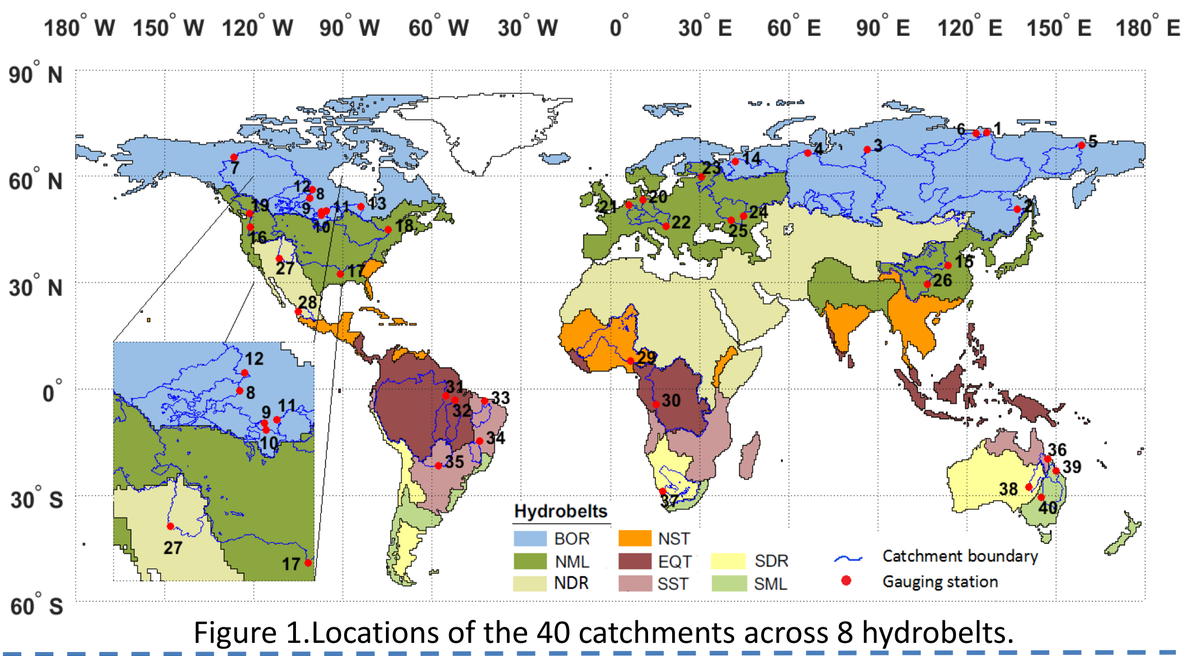


Figure 1.Locations of the 40 catchments across 8 hydrobelts.

2.3. Aggregated hydrological indicators

Six indicators of mean and extreme runoff are calculated (indicators of return period not presented):

- Mean annual runoff, MAR.
- Mean monthly runoff, MMR.
- Q5, the magnitude of monthly runoff exceeded 5 % of the time.
- Q95, the magnitude of monthly runoff exceeded 95 % of the time.

2.2. Ideal point error (IPE) for evaluating models (GHMs & EM) performance

As an integrated metric, IPE combines single metrics into one:

$$\mathrm{IPE} = [0.333 * (\left(\frac{\mathrm{RMSE_m}}{\mathrm{RMSE_b}}\right)^2 + \left(\frac{\mathrm{MARE_m}}{\mathrm{MARE_b}}\right)^2 + \left(\frac{\mathrm{CE_m} - 1}{\mathrm{CE_b} - 1}\right)^2)]^{\frac{1}{2}}$$

RMSE: Root mean squared error
MARE: Mean absolute relative error
CE: Coefficient of efficiency
***i*:** i*th* participating model (GHMs)
***max (x) or min (x)*:** the max or min value of the statistic *x* among the group of models
b: denotes metrics of a benchmark model against which the model's performance is calculated

The benchmark model is to make IPE comparable across catchments and for runoff at time step t is predicted by observed runoff at time step t-1. IPE ranges between (-∞, 1] and [1, +∞).

2.4. Weighted performance measures and performance ranking

Measures of performance are aggregated for an entire hydrobelt by calculating a weighted mean, to resolve spatial biases introduced by having different number of catchments in each hydrobelt. For each catchment, observed mean annual runoff (MAR) is applied as the relative weight, so any weighted metric (*W_m*) can be calculated as:

$$\mathrm{W\_m_{HB}} = \frac{\sum_{c=1}^{n} \mathrm{MAR_c}.\mathrm{M_c}}{\sum_{c=1}^{n} \mathrm{MAR_c}}$$

m: metric, **HB:** hydrobelt, **c:** catchment and ***n:*** number of catchments in each hydrobelt

4. Results

- Model performance is generally better in the EQT and N hydrobelts than the S hydrobelts
- The relatively lower performance in southern hemisphere hydrobelts is the result of periods of very low (or zero) runoff disproportionately effecting the IPE through inflated MARE values
- The EM outperforms the best GHM only in 2 catchments
- There is a general trend towards the over-estimation of MAR, Q95 and Q5 by all models (Fig 3)

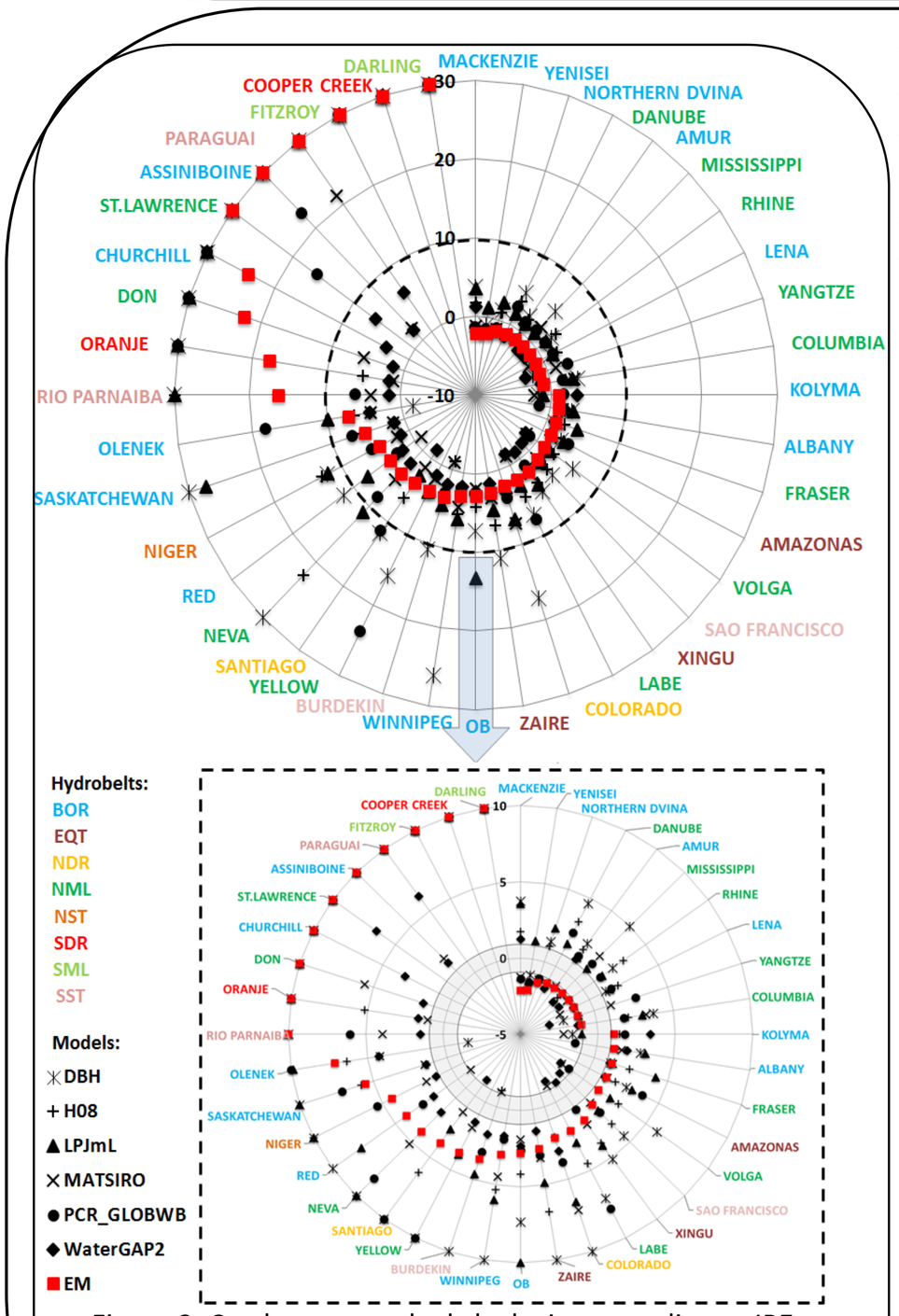
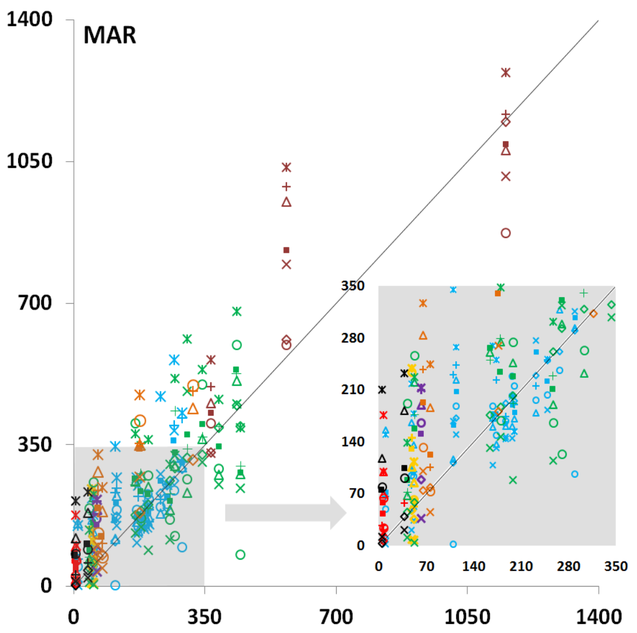
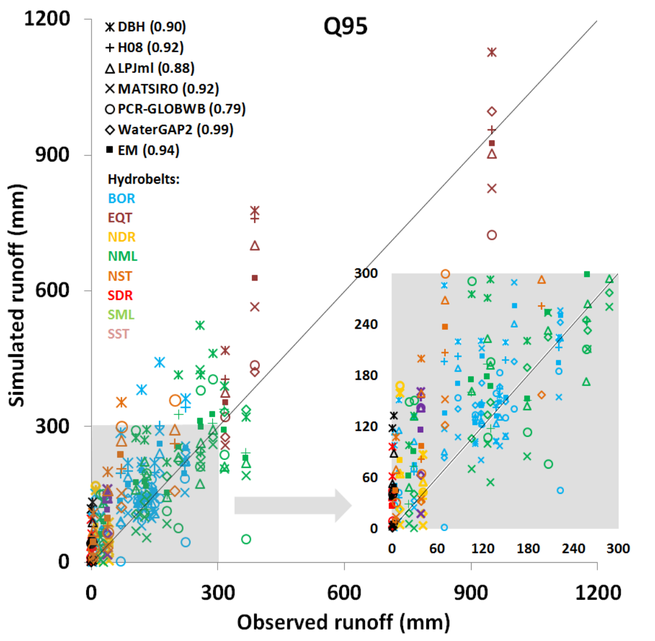


Figure 2. Catchments ranked clockwise according to IPE for the EM. IPE is capped at 30 (top panel). The bottom panel focuses on IPE≤10 with the range (-1,1) in grey, representing the boundary of performance improvements (≤-1) or loss (≥1) relative to the benchmark model.





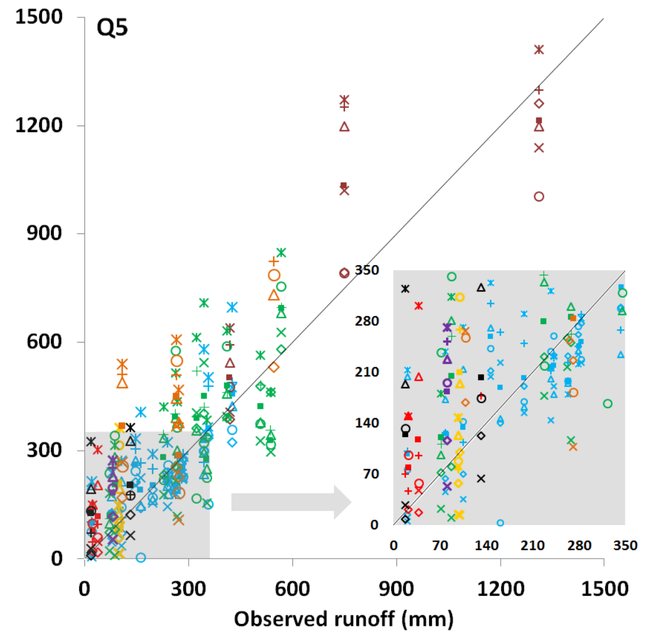


Figure 3. Scatter plots of observed vs. simulated runoff for MAR, Q5 and Q95.

Table 1. Median W-IPE by hydrobelt. Negative values indicate performance gain over the benchmark model.

Hydrobelt (#Catchs)	DBH	H08	LPJmL	MATSIRO	PCR-GLOBWB	WaterGAP2	EM	Hbelt Rank (based EM)
BOR (14)	25.6	5.3	8.1	2.5	3.5	1.1	2.0	2
NML (12)	53.0	12.7	9.8	7.2	5.8	-0.3	3.9	4
NDR (2)	14.5	4.0	4.7	4.3	7.4	0.9	2.9	3
NST (1)	12.1	12.8	12.1	1.8	5.3	1.1	4.3	5
EQT (3)	4.8	3.2	3.6	1.9	2.9	0.5	1.6	1
SST (4)	26.9	21.8	19.5	2.5	13.4	1.3	10.5	6
SDR (2)	5305.1	96.4	2051.5	326.0	520.6	78.1	1393.7	8
SML (2)	2780.7	109.5	1440.1	128.5	958.5	42.1	909.9	7
Median	26.2	12.8	10.9	3.4	6.6	1.1	4.1	
Model rank	6	5	4	2	3	1		